Skysong Innovations
Advancing the Arizona State University Knowledge Enterprise

1475 N. Scottsdale Road, Suite 200
Scottsdale, AZ 85287-3538
Phone: 480 884 1996 Fax: 480 884 1984

Case ID:M21-149P
Published: 10/12/2022

Inventors

Jae-Sun Seo
Deepak Kadetotad
Chaitali Chakrabarti
Visar Berisha

Contact

Shen Yan
shen.yan@skysonginnovations.com

Hierarchical Coarse-Grain Sparsity for Deep Neural Networks

-Background

Recurrent neural networks (RNNs) that enable accurate automatic speech recognition (ASR) are large in size and have long short-term memory (LSTM) capabilities. Due to the large size of these networks, most speech recognition tasks are performed in the cloud servers, which requires constant internet connection, introduces privacy concerns, and incurs latency for speech recognition tasks.

The particular challenge of performing on-device ASR is that state-of-the-art LSTM-based models contain tens of millions of weights. Weights can be stored on-chip (e.g., static random-access memory (SRAM) cache of mobile processors), which has the fastest access time in the nanoseconds range, but is limited to a few megabytes due to cost. Alternatively, weights can be stored off-chip (e.g., dynamic random-access memory (DRAM)) up to a few gigabytes, but access is slower in the tens of nanoseconds range and consumes nearly 100 times higher energy than on-chip counterparts. To improve energy-efficiency of neural network hardware, off-chip memory access and communication need to be minimized, and it is crucial to store most or all weights on-chip through sparsity/compression and/or weight quantization.

Invention Description

Researchers at Arizona State University have developed hierarchical coarse-grain sparsity (HCGS), a novel algorithm-hardware co-optimized memory compression technique designed to compress deep neural networks in a hardware-efficient manner. This technique allows for LSTM recurrent neural networks with up to 16 times fewer weighs while achieving minimal error rate degradation. The prototype of this design achieves up to 8.93 Tera-operations per second per watt (TOPS/W) for real-time speech recognition using compressed LSTMs based on HCGS. This technique has demonstrated energy-efficient speech recognition with low error rates for TIMIT, TED-LIUM, and LibriSpeech datasets.

Potential Applications

- Automatic speech recognition
- Machine translation
- Image recognition

Benefits & Advantages

- Maintains coarse-grain sparsity while allowing fine-grain weight connectivity (significant energy and area reduction)
- Decreased error rates

Related Publication: An 8.93 TOPS/W LSTM Recurrent Neural Network Accelerator Featuring Hierarchical Coarse-Grain Sparsity for On-Device Speech Recognition

Related Publication: Compressing LSTM networks with hierarchical coarse-grain sparsity